BIOLOGÍA EN AGRONOMÍA



Volumen 1, No. 2 **Octubre de 2011**
ISSN 1853-5216

EL USO DE INFOSTAT EN LA EXPLORACION DE DATOS BIOLOGICOS

Ovejero, Diana; Rojas, Ilda Rosa.

Facultad de Ciencias Agrarias. Universidad Nacional de Catamarca. E-mails: dianaove@yahoo.com.ar, ildarojas@argentina.com.ar

Recibido: 25/04/2011 Aceptado: 05/08/2011

RESUMEN

En la actualidad todo trabajo de investigación en ciencias aplicadas requiere de un análisis estadístico. El Análisis Exploratorio de Datos, rutina obligatoria de todo estudio científico, permite al investigador un entendimiento básico de la información obtenida y de las relaciones existentes entre las variables analizadas ya que proporciona métodos sencillos para organizar y preparar los datos, detectar fallas en el diseño y recolección de los mismos, identificar casos atípicos y comprobar los supuestos subyacentes en la mayor parte de las técnicas multivariadas. Exige un constante uso de visualizaciones gráficas y sus métodos garantizan que valores de datos extraños no influyan indebidamente en los resultados del análisis. InfoStat es un software estadístico de aplicación general desarrollado por docentes-investigadores de Estadística y Biometría y de Diseño de Experimentos de la Facultad de Ciencias Agropecuarias de la UNC. Cubre tanto las necesidades elementales para la obtención de estadísticas descriptivas y gráficos para el análisis exploratorio, como métodos avanzados de modelación estadística y análisis multivariado. Una de sus fortalezas es la sencillez de su interfaz combinada con capacidades profesionales para el cálculo y el manejo de datos. Debido al origen universitario, el programa tiene muchas facilidades para la enseñanza de la estadística que no son fáciles de encontrar en otros programas similares. El objetivo de este taller es llevar a cabo análisis

exploratorio de datos biológicos reales mediante la aplicación del Software Estadístico InfoStat.

PALABRAS CLAVES: InfoStat; Análisis exploratorio; Datos biológicos.

INFOSTAT USE FOR BIOLOGICAL DATA EXPLORATION

SUMMARY

Every research work in applied sciences requires statistical analyses. Data Exploration Analysis, a routine for all scientific work, gives the researcher a basic understanding of the information and of the relations existing between the variables analyzed because of the simple methods used to organize and prepare data, to detect flaws in the design and collection, to identify atypical cases, and to prove the underlying suppositions in most multivariate techniques. Data Exploration Analysis demands constant use of graphic displays, and the methodology ensures that strange data values do not influence improperly the results of the analysis. InfoStat is a statistical software of general application designed by teachers-researchers of Statistics and Biometry, and of Experiments Design of the Agricultural Sciences College at the National University of Cordoba. InfoStat meets the demands to obtain both descriptive statistics and graphics for exploration analysis, and advanced methods of statistical modelling and multivariate analysis. One of the strong points of this software is the simplicity of the interface combined with professional capacities for calculation and data management. Because this program originated in a university, it is simpler for Statistics teaching than other similar programs. The objective of this workshop is to develop exploration analysis of true biological data by means of the application of the Statistics software InfoStat.

KEY WORDS: InfoStat; Exploration analysis; Biological data.

FUNDAMENTACIÓN

En la actualidad todo trabajo de investigación en ciencias aplicadas requiere de un análisis estadístico. El Análisis Exploratorio de Datos, rutina obligatoria de todo

estudio científico, permite al investigador un entendimiento básico de la información obtenida y de las relaciones existentes entre las variables analizadas ya que proporciona métodos sencillos para organizar y preparar los datos, detectar fallas en el diseño y recolección de los mismos, identificar casos atípicos y comprobar los supuestos subyacentes en la mayor parte de las técnicas multivariadas

La finalidad del Análisis Exploratorio de Datos es examinar los datos previamente a la aplicación de cualquier técnica estadística.

OBJETIVOS

- Realizar un análisis exploratorio de datos biológicos reales mediante el uso del Software Estadístico InfoStat.

CONTENIDOS

Análisis Exploratorio de Datos. Etapas del Análisis Exploratorio de Datos. Preparación de los Datos. Análisis Estadístico Unidimensional. Datos Atípicos (outliers). Datos Ausentes (missing).

- **Análisis exploratorio de datos**

El Análisis Exploratorio de Datos, se utiliza en las fases iniciales de un estudio experimental, y consiste en el estudio de los datos desde todas las perspectivas, y con todas las herramientas posibles. Exige un constante uso de visualizaciones gráficas y sus métodos garantizan que valores de datos extraños no influyan indebidamente en los resultados del análisis.

- **Etapas del análisis exploratorio de datos**

a) Preparar los datos para hacerlos accesibles a cualquier técnica estadística.

b) Realizar un gráfico de las variables individuales a analizar y un análisis descriptivo.

c) Analizar las relaciones entre las variables para ver el grado de interrelación existente entre ellas.

d) Identificar los casos atípicos (outliers) y evaluar el impacto que puedan ejercer en análisis estadísticos posteriores.

6) Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

- **Preparación de los datos**

En un **Análisis Exploratorio de Datos** se hace necesario preparar los datos para aplicar cualquier técnica estadística. Esto implica la forma en que serán introducidos los datos, la codificación, en caso necesario (depende del tipo de variable), como así también, la selección de un paquete estadístico adecuado para procesarlos.

Los paquetes estadísticos son conjuntos de programas que implementan diversas técnicas estadísticas en un entorno común. Los paquetes estadísticos son conjuntos de programas que implementan diversas técnicas estadísticas en un entorno común. Entre los más utilizados están SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Sciences), STATA, S-PLUS, R, STATGRAPHICS, InfoStat.

- **Análisis estadístico unidimensional**

Una vez organizados los datos, se realiza un análisis estadístico gráfico y numérico de las variables del problema con el fin de tener una idea inicial de la información contenida en el conjunto de datos, así como detectar la existencia de posibles errores en la codificación de los mismos. El análisis a realizar depende de la escala de medida de la variable a analizar.

- **Datos Atípicos (outliers)**

Los datos atípicos son observaciones con características diferentes de las demás.

Estos casos deben ser contemplados en el contexto del análisis y es necesario evaluar el tipo de la información que proporcionar. Pueden ser elementos no representativos de la población y por lo tanto encubrir (distorsionar) el comportamiento de los datos. En ocasiones son elementos no representativos de la población y en consecuencia distorsionan el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.

- **Datos Ausentes (missing)**

Los datos ausentes son algo habitual en el Análisis Multivariante; rara es la investigación en la que no aparece este tipo de datos.

En estos casos el investigador debe tratar de averiguar las razones que subyacen en el dato ausente buscando entender el proceso principal de esta ausencia para seleccionar el curso de acción más apropiado. Los datos ausentes pueden deberse a errores externos al encuestado o por acción del mismo tal como rehusarse a contestar.

El investigador debe analizar si existe algún patrón no aleatorio en dicho proceso que pueda sesgar los resultados obtenidos debido a la pérdida de representatividad de la muestra analizada.

MATERIAL Y MÉTODO

Origen de los datos

En el segundo cuatrimestre de 2007 los alumnos de segundo año de la Carrera Ingeniería Agronómica llevaron a cabo el “Taller de intensificación de la actividad práctica” Las cátedras participantes fueron: Biometría y Técnica Experimental, Botánica II, Física II y Química Analítica. Dicho taller se llevó a cabo en un establecimiento olivícola del Valle Central de la Pcia. de Catamarca ubicado en la localidad de Las Esquinas, Valle Viejo . Se trabajó sobre dos lotes de monovarietales de Coratina y Arbequina respectivamente, de 8 años de edad en un marco de plantación de 5x4. Cada lote contaba con 32 filas con 39 planta por fila

El suelo es franco arenoso levemente alcalino, con valores intermedios de materia orgánica, bajos en N y adecuados de P y K.

El 19 de septiembre de 2007 se seleccionaron, de cada lote 25 plantas y se les midió el diámetro (cm) con cinta métrica a una altura de 10 cm del suelo. De cada planta se seleccionaron 4 brindillas a altura del operador (una por cada punto cardinal) y se colocaron en bolsas de papel con datos identificatorios.

Mediciones en laboratorio de Física

En el laboratorio de Física se consignaron, para cada variedad, los datos correspondientes a las siguientes variables:

- Longitud de brindillas (cm): desde la base hasta la última hoja extendida
- Numero de nudos por brindilla
- Densidad de nudos por brindilla
- Cantidad de flores por brindilla

- Cantidad de brotes por brindilla
- Cantidad de inflorescencias en estado C,D,DI,DII (Según De Andrés).

De cada brindilla se seleccionaron 6 hojas (2 basales, 2 medias, 2 apicales) se les midió la longitud y la latitud.

Análisis en el laboratorio de Química

Se determinó el fósforo foliar y el potasio foliar de cada una de las muestras seleccionadas (clasificadas por variedad). La unidad de medida fue ppm.

Variables a utilizar

- Variedad (variable cualitativa, factor de clasificación)
- Longitud de brindillas: de base a yema apical (variable continua, variable respuesta)
- Numero de nudos por brindilla (variable discreta, respuesta)
- Cantidad de flores por brindilla (variable discreta a transformar en categórica)

Software Estadístico InfoStat

InfoStat ofrece distintas herramientas para que el usuario pueda explorar su información de manera muy sencilla, para ello trabaja con tres tipos de ventanas: la ventana donde se encuentran los datos (**Datos**), aquella donde se muestran y acumulan los resultados de los procedimientos solicitados (**Resultados**) y la ventana donde se muestran y acumulan los gráficos realizados por el usuario (**Gráficos**). Las ventanas **Resultados** y **Gráficos** contienen una hoja para cada resultado y/o gráfico producido.

A través de menú ESTADÍSTICAS InfoStat ofrece la posibilidad de obtener de manera casi automática estadística descriptiva, calcular probabilidades, estimar características poblacionales bajo distintos planes de muestreo, estadística inferencial para una y dos muestras mediante diversos tipos de intervalos de confianza y pruebas de hipótesis (paramétrica y no paramétrica), utilizar modelos de regresión y análisis de varianza para distintos tipos de experimentos diseñados y estudios observacionales, estadística inferencial para datos categorizados, entre otras muchas posibilidades. estadística multivariada, análisis de series de tiempo, suavizados y ajustes.

El diseño de investigación a aplicar es descriptivo, pues se realiza un análisis exploratorio de los datos.

Estudio comparativo entre variedades

Se aplicarán técnicas elementales del análisis exploratorio de datos, como ser**:** tablas de frecuencias, gráficos de frecuencias, gráfico de cajas y medidas de resumen.

Tablas de frecuencias: es una tabla en la que se organizan los datos obtenidos en grupos de valores y además, muestra la cantidad de observaciones del conjunto de datos que caen en cada uno de esos grupos.

Gráficos de frecuencias: la representación gráfica de una tabla de frecuencias depende del tipo de datos con el que se trabaje.

Gráfico de cajas y patillas: este gráfico proporciona, la posición relativa de la mediana, cuartiles y extremos de la distribución. Además, proporciona información sobre los valores atípicos, e informa de la simetría o asimetría de la distribución.

El gráfico de la caja también se puede utilizar para comparar la misma variable en muestras distintas.

Medidas de resumen: las tablas de frecuencias y sus representaciones gráficas son útiles a los fines descriptivos sin embargo, cuando la variable es cuantitativa estas descripciones pueden ser aún, poco prácticas a los fines comparativos. Por ello se utilizan medidas de resumen que caracterizan a estas distribuciones**.**

Las medidas de resumen aportan la información acerca de **valores centrales, la dispersión y la forma de la distribución.**

Variable: Longitud brindillas

Tabla 1: Longitud de Brindillas - Variedad: Coratina

Longitud (cm)		Brindillas		
LI	LS	Cantidad	%	% Acumulado
16,00	26,00	4	4	4
26,00	36,00	26	21	30
36,00	46,00	34	35	64
46,00	56,00	24	24	88
56,00	66,00	8	11	96
66,00	76,00	4	5	100
Total		100	100	

Distribución asimétrica a derecha. El 84% de las brindillas poseen longitudes entre 26 y 56 cm. El 12%, longitudes mayores o iguales a 56cm.

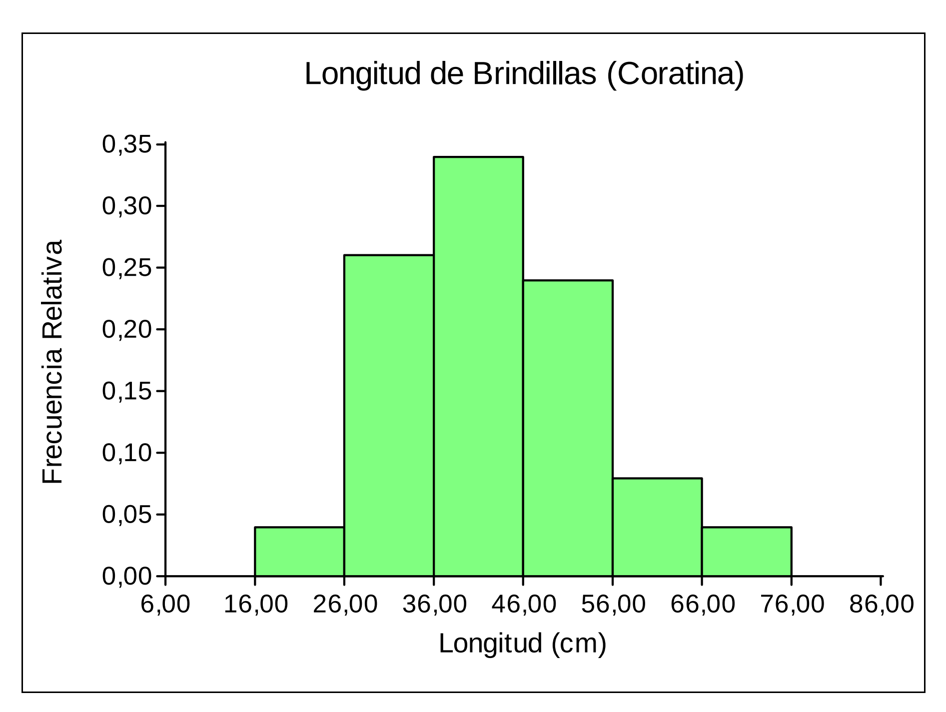


Gráfico 1: Longitud de brindillas (Coratina)

Tabla 2: Longitud de Brindillas - Variedad: Arbequina

Longitud (cm)		Brindillas		
LI	LS	Cantidad	%	% Acumulado
19,00	25,50	15	15	15
25,50	32,00	22	22	37
32,00	38,50	26	26	63
38,50	45,00	25	25	88
45,00	51,50	9	9	97
51,50	58,00	3	3	100
Total		100	100	

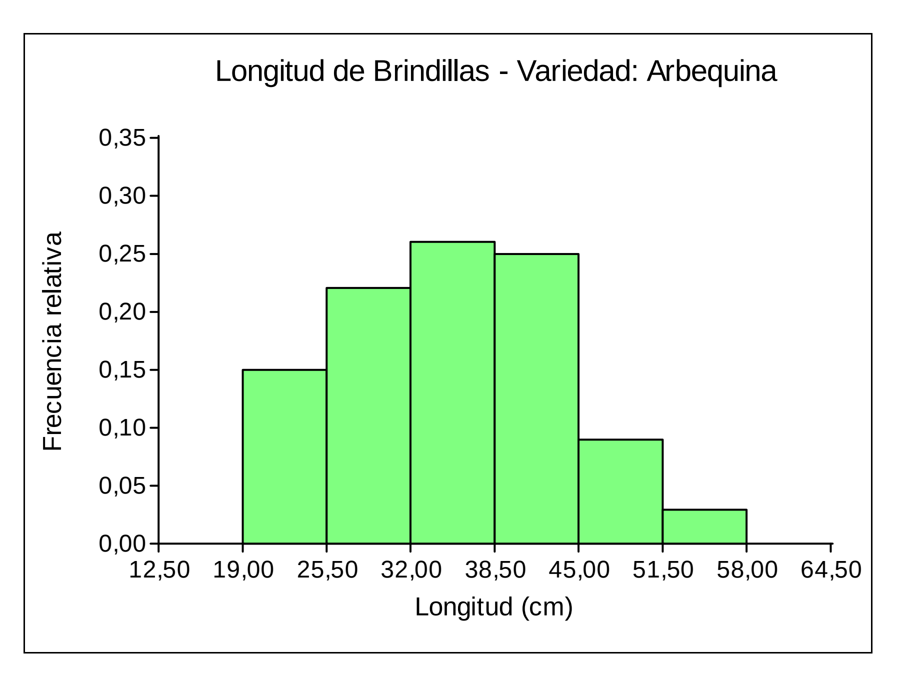


Gráfico 2: Longitud de brindillas (Arvequina)

Al igual que en la variedad Coratina la distribución es asimétrica a derecha. El 73% de las brindillas tiene longitudes entre 25,5 y 45 cm. No hay brindillas con longitudes superiores o iguales a 58 cm. ni menores a 19 cm . El 12% tiene longitudes iguales o mayores a 45 cm.

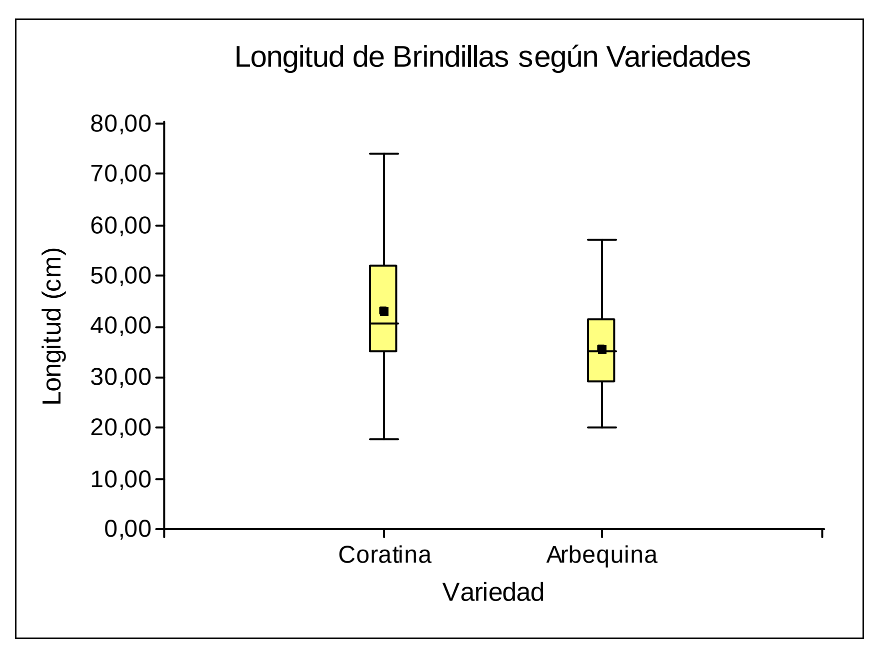


Gráfico3: Longitud de brindillas según variedades

Tabla 3: Longitud de Brindillas - Medidas de Resumen según Variedades

Resumen	Variedades	
	Coratina	Arbequina
n	100,000	100,000
Media	43,012	35,340
D.E.	11,530	8,414
Var(n-1)	132,949	70,080
CV	26,807	23,810
Mín	17,800	20,000
Máx	74,000	57,000
Mediana	40,750	35,000
Q1	34,600	28,900
Q3	52,200	41,500

Las longitudes de la variedad Coratina son más variables. DE Coratina = 11,53 Vs DE Arbequina = 8,414

El rango intercuartílico en Coratina es de 17,6 mientras que el de Arbequina es de 15,6 (El 50% central de las longitudes de Arbequina está más concentrado que el correspondiente en Coratina). La mediana de las longitudes de Arbequina (35,00) posee un valor próximo al primer cuartil de Coratina (34,600). Es decir que el 50% de las longitudes de Arbequina poseen longitudes mayores o iguales a 35cm mientras que el 75% de las longitudes de Coratina tienen longitudes mayores o iguales a 34,6cm.

Variable: Cantidad de nudos por brindilla

Tabla 4: Cantidad de Nudos por Brindilla - Variedad: Coratina

Cantidad de Nudos	Brindillas			
	Cantidad	%	Cantidad Acumulada	% Acumulado
8	2	2	2	2
10	2	2	4	4
11	3	3	7	7
12	2	2	9	9
13	11	11	20	20
14	8	8	28	28
15	11	11	39	39
16	6	6	45	45
17	11	11	56	56
18	14	14	70	70
19	10	10	80	80
20	9	9	89	89
21	3	3	92	92
22	3	3	95	95
23	3	3	98	98
24	1	1	99	99
25	1	1	100	100
Total	100	100		

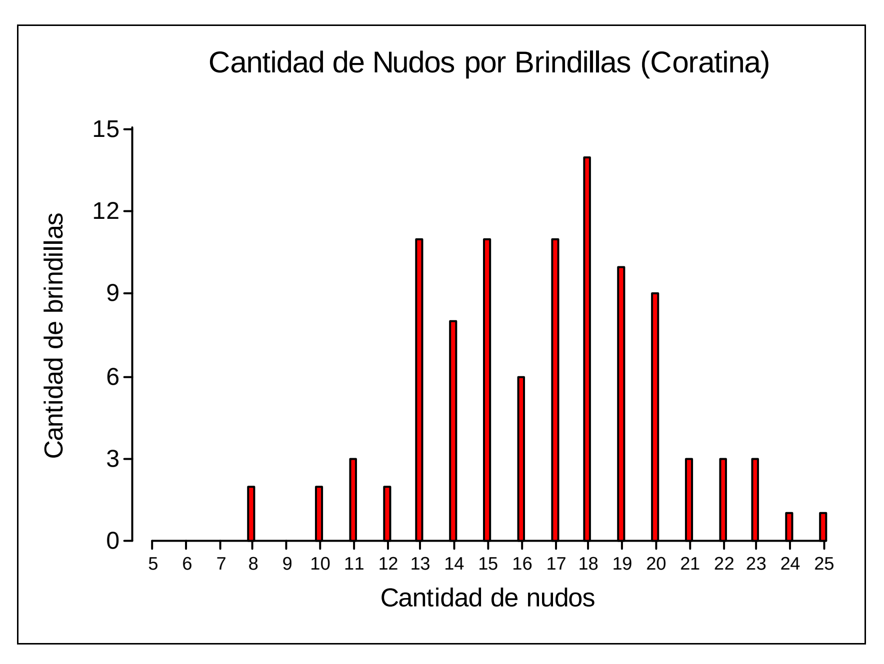


Gráfico 4: Cantidad de nudos por brindillas (Coratina)

Distribución aproximadamente simétrica. Entre 13 y 19 nudos por brindilla se encuentra el 71% de la muestra. El 11% de las brindillas tiene más de 20 nudos.

Tabla 5: Cantidad de Nudos por Brindilla - Variedad: Arbequina

Cantidad de Nudos	Brindillas			
	Cantidad	%	Cantidad Acumulada	% Acumulado
10	1	1	1	1
13	1	1	2	2
14	1	1	3	3
15	2	2	5	5
16	1	1	6	6
17	4	4	10	10
18	1	1	11	11
19	6	6	17	17
20	8	8	25	25
21	10	10	35	35
22	10	10	45	45
23	5	5	50	50
24	11	11	61	61
25	6	6	67	67
26	6	6	73	73
27	5	5	78	78
28	4	4	82	82
29	5	5	87	87
30	3	3	90	90
31	3	3	93	93
32	3	3	96	96
33	4	4	100	100
Total	100	100		

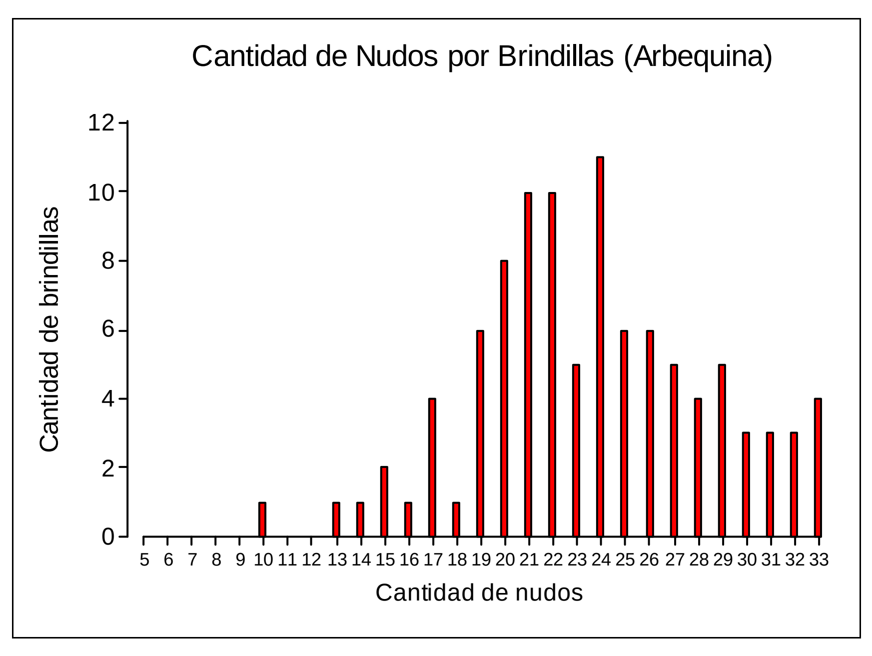


Gráfico 5: Cantidad de nudos por brindilla (Arbequina)

Distribución asimétrica a izquierda. Entre 19 y 28 se encuentra el 71% de las brindillas

El 39% de las brindilla tiene más de 24 nudos por brindilla

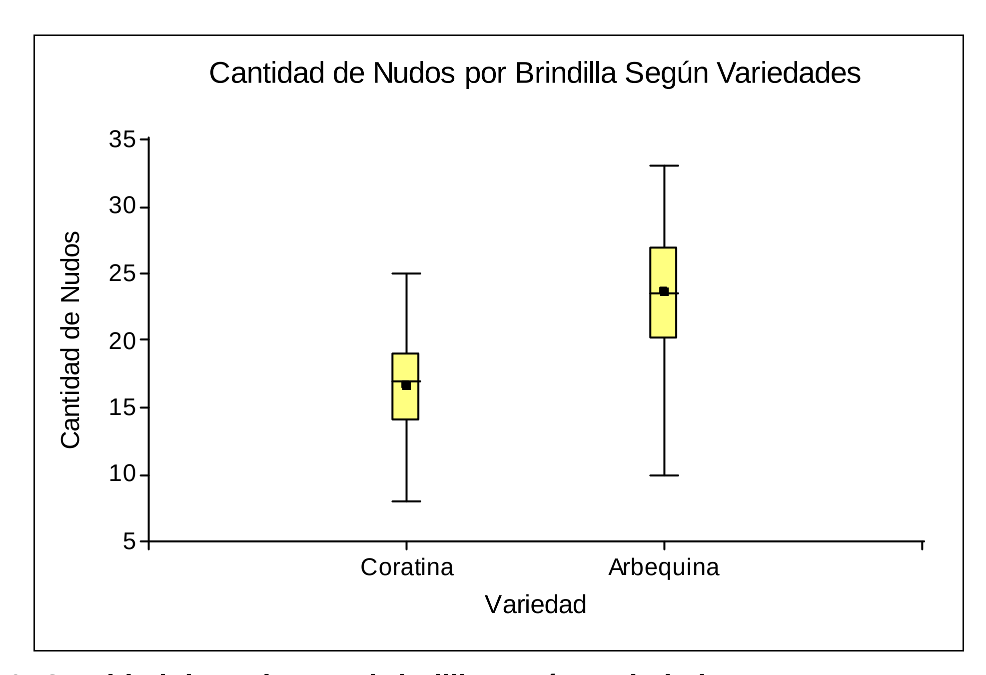


Gráfico 6: Cantidad de nudos por brindilla según variedades

Tabla 6: Cantidad de Nudos por Brindilla - Medidas de Resumen según Variedades

Resumen	Variedades	
	Coratina	Arbequina
n	100,00	100,00
Media	16,65	23,61
D.E.	3,46	4,83
Var(n-1)	11,95	23,31
CV	20,76	20,45
Mín	8,00	10,00
Máx	25,00	33,00
Mediana	17,00	23,50
Q1	14,00	20,00
Q3	19,00	27,00

La cantidad de nudos por brindilla en la variedad Arbequina es más variables. (DE Arbequina = 4,83 Vs DE Coratina = 3,46).

Es marcada la diferencia entre el número promedio de nudos por brindilla entre las dos variedades (Coratina = 16,65 Vs Arbequina = 23,61).

El tercer cuartil de Coratina (19) es menor que el primer cuartil de Arbequina (20) es decir que en Coratina el 75% de las brindillas tiene 19 ó menos nudos mientras que en Arbequina sólo el 25% de las brindillas tiene 20 ó menos nudos.

En Coratina no hay brindillas con más de 25 nudos, en Arbequina el 25% de las brindillas posee 27 ó más nudos

Variable: Cantidad de flores por brindilla

Se dicotomiza la variable y se obtiene

Tabla 7: Presencia de Flores en las Brindillas – Variedad: Coratina

Presencia de Flores	Brindillas	
	Cantidad	%
No	97	0,97
Sí	3	0,03
Total	100	100

Tabla 8: Presencia de Flores en las Brindillas – Variedad: Arbequina

Presencia de Flores	Brindillas	
	FA	FR
No	2	0,02
Sí	98	0,98
Total	100	100

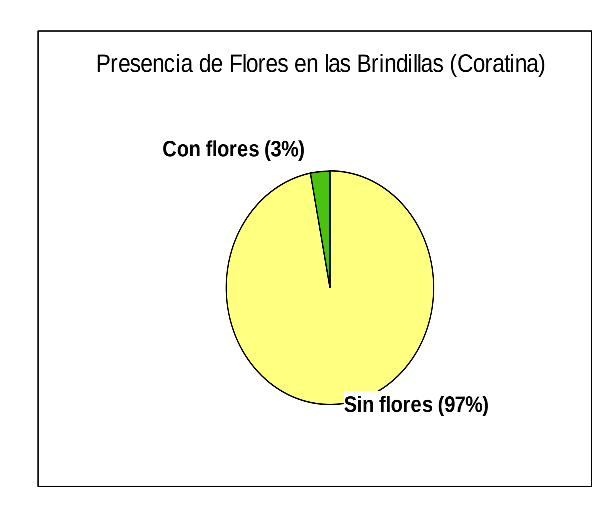


Gráfico 7: Presencia de flores en las brindillas (Coratina)

Gráfico 8: Presencia de flores en las brindillas (Arbequina)

Resulta evidente la diferencia que existe entre las dos variedades respecto de la presencia de flores en las brindillas. En Coratina únicamente el 3% de las brindillas presentaban flores, mientras que en Arbequina el 98% de las bindillas tienen flores.

EVALUACIÓN

Realizar un análisis exploratorio de datos tomando una base de no menos de 50 datos.

AREAS DE INTERES

Biologia, Agronomia, Veterinaria.

BIBLIOGRAFIA

- Balzarini M.G., Gonzalez L., Tablada M., Casanoves F., Di Rienzo J.A., Robledo C.W. (2008). *Manual del Usuario*, Editorial Brujas, Córdoba, Argentina.
- Salvador Figueras, M y Gargallo, P. (2003): "Análisis Exploratorio de Datos".
- Batanero, C.; Estepa, A. y J. D. Godino (1991) “Análisis Exploratorio de Datos: sus posibilidades en la enseñanza secundaria”.
- Tablada Elena Margot, Córdoba, Mariano, Balzarini Mónica (2011) “Análisis exploratorio de datos”. *http://agro.uncor.edu/~estad/*SOFTWARE

- Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W. InfoStat versión 2011. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

DIRECCIONES Y ENLACES DE INTERÉS

- Batanero, C., Estepa, A. y Godino, J. D. (1991). Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria . *Suma,* 9, 25-31 http://www.ugr.es/~batanero/publicaciones%20index.htm
- Batanero, C. (1999). Taller sobre analisis exploratorio de datos en la enseñanza secundaria . *Actas de la Conferencia Internacional "Experiências e Expectativas do Ensino de Estatística - Desafios para o Século XXI"* . Florianópolis, Santa Catarina, Brasil - 20 a 23 de Setembro de 1999. http://www.ugr.es/~batanero/publicaciones%20index.htm
- Tablada, M.; Córdoba, M.; Balzarini, M. (2011). Análisis exploratorio de datos http://agro.uncor.edu/~estad/
- Salvador Figueras, M y Gargallo, P. (2003): "Análisis Exploratorio de Datos", http://www.5campus.com/leccion/aed
- Para descargar InfoStat Estudiantil http://www.infostat.com.ar/